STATISTICS 110, FALL 2015 FINAL EXAM

NAME:________KEY______________________ Your assigned homework number:__________

Last 6 digits of Student ID: _________________ Assigned seat for this exam: ______________

Open notes. You should have 6 pages plus a page of output. Use the back of the pages if you need more space. Each problem is worth **6 points** except where indicated.

1. (2 pts each) In a linear regression situation with response variable *Y* and one or more *X* explanatory variables, specify whether each of the following involves the *Y* values only, the *X* values only, or both the *Y* and the *X* values. *Circle your answer.*

a. Variance inflation factor for X_1	Ys only	***Xs only***	Ys and Xs
b. Hat values	Ys only	***Xs only***	Ys and Xs
c. SSTotal	***Ys only***	Xs only	Ys and Xs
d. Cook's Distance values	Ys only	Xs only	***Ys and Xs***
e. Predicted values ($\hat{Y}$)	Ys only	Xs only	***Ys and Xs***

The following scenario is for questions 2 to 10:
A pharmaceutical company is developing a new drug that it hopes will provide relief for hay fever. They are considering two active ingredients, which we will call A and B. For the initial part of the experiment they decided to test only ingredient A, to figure out what concentration of that ingredient works best. They create identical-looking pills but that have four different concentration levels, including a placebo that has none of the active ingredient. The concentration levels are 0 mg (placebo), 10 mg, 15 mg and 20 mg. 100 volunteers who suffer from hay fever are willing to participate in the experiment. 25 volunteers are randomly assigned to each of the concentration levels. The response variable is a "relief score" found by using number of hours of relief, and subtracting points for negative side effects. High relief scores are desirable. A plot and some output are given on a separate page.

2. Two possible analysis methods are simple linear regression with X = concentration level, or one-factor ANOVA, with the concentration levels treated as a categorical variable with 4 categories. The output page shows summary statistics and a plot of the results with Y = relief score and X = concentration level. Explain why it would not be appropriate to use simple linear regression as the analysis method.

   One of the necessary conditions for simple linear regression is that the relationship between Y and X is approximately linear. It is clear from the plot that this condition is not met.

3. The analysts decide to use one factor ANOVA. They define Y_{ik} = relief score for person *i* taking concentration *k*, with *k* = 1, 2, 3, 4, for placebo, 10 mg, 15 mg and 20 mg, respectively, and *i* = 1 to 25 for each *k*. They define 4 indicator variables, with Ak = 1 if the individual took the pill with concentration level k, and 0 otherwise. They used the following model, omitting A1*,* the indicator variable for the placebo group: $Y = \beta_0 + \beta_1 A2 + \beta_2 A3 + \beta_3 A4 + \varepsilon$

   a. (3 pts) Interpret the coefficient β_0 in this situation.

   β_0 is the population mean relief score for the placebo condition. "Population mean" in this case would be the mean for all people similar to the ones in this experiment, if they were to take the placebo.

   b. (3 pts) Interpret the coefficient β_1 in this situation.

   β_1 is the difference between the population mean relief score if everyone in the population were to take 10mg versus if they were to take the placebo (10 mg mean – placebo mean).

   c. (3 pts each) Using information provided on the page of output, give the values of $\hat{\beta}_0$ and $\hat{\beta}_1$

   $\hat{\beta}_0$ = _7.5405______ $\hat{\beta}_1$ = _14.7324 – 7.5404 = 7.192__

4. Results for the Tukey method are shown on the output. The 6 possible pairs of means are listed below. Based on the Tukey results, which means are significantly different using family α = 0.05? Circle Yes if they are significantly different, and No if they are not.

   Placebo and 10 mg? ***Yes*** No   10 mg and 15 mg? ***Yes*** No

   Placebo and 15 mg? Yes ***No***   10 mg and 20 mg? Yes ***No***

   Placebo and 20 mg? ***Yes*** No   15 mg and 20 mg? ***Yes*** No

5. Based on these results, what concentration level(s) of ingredient A would you recommend the company use in their pills? (If you think there is more than one acceptable concentration level, give them all.) Explain your answer.

   Either 10 mg or 20 mg. Both produced mean relief scores significantly higher than placebo and 15 mg, but not significantly different from each other.

Additional information for Questions 6 to 10: The company decided to continue the experiment by making new pills that contained both ingredient A and ingredient B. They used 2 levels of each ingredient – none or 10 mg, so there were 4 combinations, with "none, none" representing an overall placebo. Again they had 100 volunteers, so they randomly assigned 25 to each of the 4 combinations.

6. (4 pts) Are blocks used in this experiment? Briefly explain.

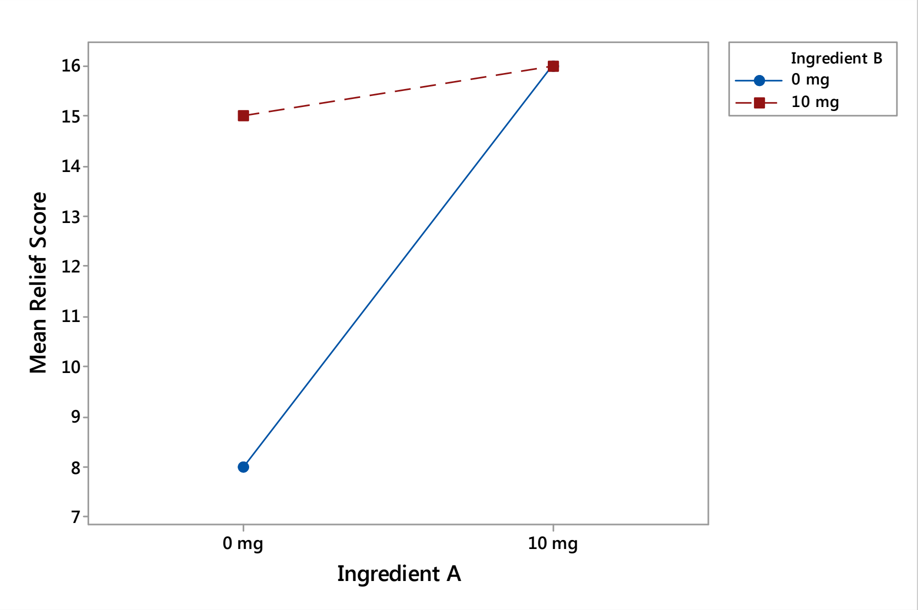
   No. Each volunteer was measured only once.

7. (8 pts) There are two factors in this experiment. Name each factor, and then specify whether it is fixed or random, how many levels it has, and what the levels are.

Factor name	Fixed or Random?	Number of levels	Levels
Ingredient A	*Fixed*	2	0, 10 mg
Ingredient B	*Fixed*	2	0, 10 mg

8. The relief score means for the combination of ingredients are shown in the table below. Use them to create an interaction plot on the axes provided. Label everything clearly.

	Concentration of Ingredient B	
Concentration of Ingredient A	**None**	**10 mg**
None	8	15
10 mg	16	16



9. Based on the cell means and your plot in Question 8, comment on whether there appears to be a non-zero Factor A effect, Factor B effect and/or interaction effect, and explain briefly how you know.

   Factor A effect?

   Yes, because the average for A_1 (none) is (8 + 15)/2 = 11.5, but the average for A_2 (10 mg) is 16.

   Factor B effect?

   Yes, because the average for B_1 (none) is (8 + 16)/2 = 12, but the average for B_2 (10 mg) is 15.5.

   Interaction effect?

   Yes. There are a few ways you can write the explanation. For instance, you can say that the change in relief score when going from none to 10 mg of ingredient A depends on how much of ingredient B is in the pill. If none of Ingredient B is in it, the change is large, jumping from 8 to 16. But if 10 mg of ingredient B is in it, the change is small, from 15 to 16.

10. In this situation, the full model can be written as $Y = \mu + \alpha_k + \beta_j + \gamma_{jk} + \varepsilon$. Using this notation, write the model corresponding to each of the following null hypotheses.

    a. H_0: There is no interaction and no Factor A effect.

    $$Y = \mu + \beta_j + \varepsilon.$$

    b. H_0: There is a Factor B effect and an interaction, but no Factor A effect.

    $$Y = \mu + \beta_j + \gamma_{jk} + \varepsilon.$$

    c. If you wanted to use the hypothesis in part (a) as the null hypothesis and the hypothesis in part (b) as the alternative hypothesis, could you use the nested F test (i.e., the full and reduced model framework)? Explain your answer.

    Yes. The model in (a) contains a subset of the terms in the model in (b).

MULTIPLE CHOICE (3 points each); circle the best answer.
The following scenario is for Questions 1 to 6:
A university would like to reduce its "carbon footprint" and would like to know what incentives might get people to use less energy. Participants can earn points by visiting a website and pledging to take certain energy-saving actions. The university would like to compare 3 plans for how people are rewarded for earning points, to see which one gets people to earn the most points. The plans are:
Plan 1: Participants can redeem points for food discounts on campus.
Plan 2: Participants can redeem points for prizes such as tee shirts with the campus logo.
Plan 3: Participants can redeem points for free tickets to campus sporting events.
The university population consists of students, staff and faculty, and knowing that the 3 cohorts might have different preferences, the experiment will be done using a random sample of 120 people from each cohort, and randomly assigned 40 in each of them to each of the 3 plans. The response variable is the number of points each person earns during one month in the program. Thus, there are two factors: Factor A is the plan assigned (1, 2, or 3) and Factor B is the person's cohort (student, staff, faculty).

1. Would the participants in this study be considered to be blocks, and why?
   A. Yes, because they were randomly selected from all possible students, staff or faculty.
   B. Yes, because they were randomly assigned to one of the plans.
   C. Yes, because different individuals were in the student, staff and faculty cohorts.
   D. No, because each participant was measured only once.

2. If there is a significant "plan effect" could the university conclude that there is a *cause and effect* relationship between the plan assigned and the points earned?
   A. Yes, because a random sample of people from each cohort was used for the study.
   B. Yes, because the participants were randomly assigned to the plans.
   C. No, because people didn't volunteer for the experiment, they were selected.
   D. No, because students might prefer one plan, while faculty or staff might prefer a different plan.

3. Could the university generalize the results of the experiment to all individuals in the populations represented by each cohort (students, staff, faculty), and why?
   A. Yes, because a random sample of people from each cohort was used for the study.
   B. Yes, because the participants were randomly assigned to the plans.
   C. No, because people didn't volunteer for the experiment, they were selected.
   D. No, because students might prefer one plan, while faculty or staff might prefer a different plan.

4. What would it mean if there was a Factor A effect in this experiment?
   A. The mean points that would be earned if everyone in the university were to participate is not the same for students, staff and faculty.
   B. The mean points that would be earned if everyone in the university were to participate is not the same for all 3 plans.
   C. The preference for one plan over another is not the same for the 3 cohorts.
   D. The mean points that would be earned if everyone in the university were to participate is greater than 0.

5. What would it mean if there was a Factor B effect in this experiment?
   A. The mean points that would be earned if everyone in the university were to participate is not the same for students, staff and faculty.
   B. The mean points that would be earned if everyone in the university were to participate is not the same for all 3 plans.
   C. The preference for one plan over another is not the same for the 3 cohorts.
   D. The mean points that would be earned if everyone in the university were to participate is greater than 0.

6. What would it mean if there was an AxB interaction effect in this experiment?
   A. The mean points that would be earned if everyone in the university were to participate is not the same for students, staff and faculty.
   B. The mean points that would be earned if everyone in the university were to participate is not the same for all 3 plans.
   C. The preference for one plan over another is not the same for the 3 cohorts.
   D. The mean points that would be earned if everyone in the university were to participate is greater than 0.

The following scenario is for Questions 7 and 8:
A multiple regression model is run in R using the lm command, and then the anova(model) command is used, resulting in the following ANOVA table:

```
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
X1         1  1.108  1.1081  0.2998 0.58978
X2         1  0.770  0.7698  0.2083 0.65281
X3         1 24.713 24.7134  6.6865 0.01724 *
Residuals 21 77.616  3.6960
```

7. Using notation SS(A|B) which of the following represents the value 0.770?
   A. SS(X_2)
   B. SS(X_1 | X_2)
   C. SS(X_2 | X_1)
   D. SS(X_2 | X_1, X_3)

8. Which of the following is the value of SSModel?
   A. 77.616 – (1.108 + 0.770 + 24.713)
   B. 1.108 + 0.770 + 24.713
   C. 1.108 + 0.770 + 24.713 + 77.616
   D. It cannot be determined from the information in the output.

9. In a multiple regression setting, which one of the following is most affected if you add an explanatory variable that's highly correlated with the ones already in the equation?
   A. The overall F test
   B. The predicted values
   C. The interpretation of MSE
   D. The interpretation of the individual coefficients

10. Consider a regression situation with a quantitative variable X, a response variable Y, and one categorical variable with 3 levels. Three indicator variables A_1, A_2 and A_3 are defined, with $A_k = 1$ if the individual is from level k, and 0 otherwise. To allow different *slopes* (for the relationship between X and Y) for each level of the categorical variable, which of the following terms need to be included in the model in addition to $\beta_0 + \beta_1 X + \varepsilon$?
    A. A_1, A_2 and A_3.
    B. A_1 and A_2, but not A_3.
    C. A_1X, A_2X and A_3X.
    D. A_1X and A_2X, but not A_3X.